CDF Farms in Run II

Stephen Wolbers

CHEP2000

February 7-11, 2000

CDF Farms Group:

Jaroslav Antos, Antonio Chan, Paoti Chang, Yen-Chu Chen, Stephen Wolbers, GP Yeh, Ping Yeh

Fermilab Computing Division:

Mark Breitung, Troy Dawson, Jim Fromm, Lisa Giacchetti, Tanya Levshina, Igor Mandrichenko, Ray Pasetes, Marilyn Schweitzer, Karen Shepelak, Dane Skow

Outline

- **Requirements for Run 2 computing**
- **Design**
- **Experience, including Mock Data Challenge 1**
- **Future -- MDC 2, Run 2, Higher rates and scaling**

Requirements

- **The CDF farms must have sufficient capacity for Run 2 Raw Data Reconstruction**
- **The farms also must provide capacity for any reprocessing needs**
- **Farms must be easy to configure and run**
- **The bookkeeping must be clear and easy to use**
- **Error handling must be excellent**

Requirements

- **Capacity**
  - **Rates : 75 Hz max (28 Hz average)**
    - **250 Kbyte input event size**
    - **60 Kbyte output event size**
  - **Translates to 20 Mbyte/s input (max) and 5 Mbyte/sec output (max)**
  - **Substantially greater than Run I but not overwhelming in modern architectures**

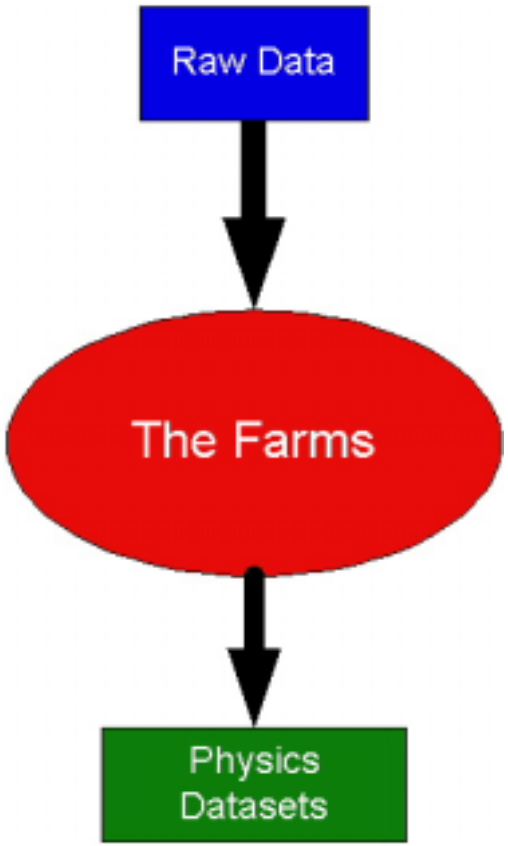
Requirements (CPU)

- **CPU goal is <5 seconds/event on PIII/500**
- **Assuming 70% efficiency this translates to**
  - **200 PIII/500 equivalents**
  - **4200 SpecInt95**
- **Adding in reprocessing, simulation, responding to peak rates**
  - **300-400 PIII/500 equivalents (150-200 duals)**
  - **6300-8400 SpecInt95**

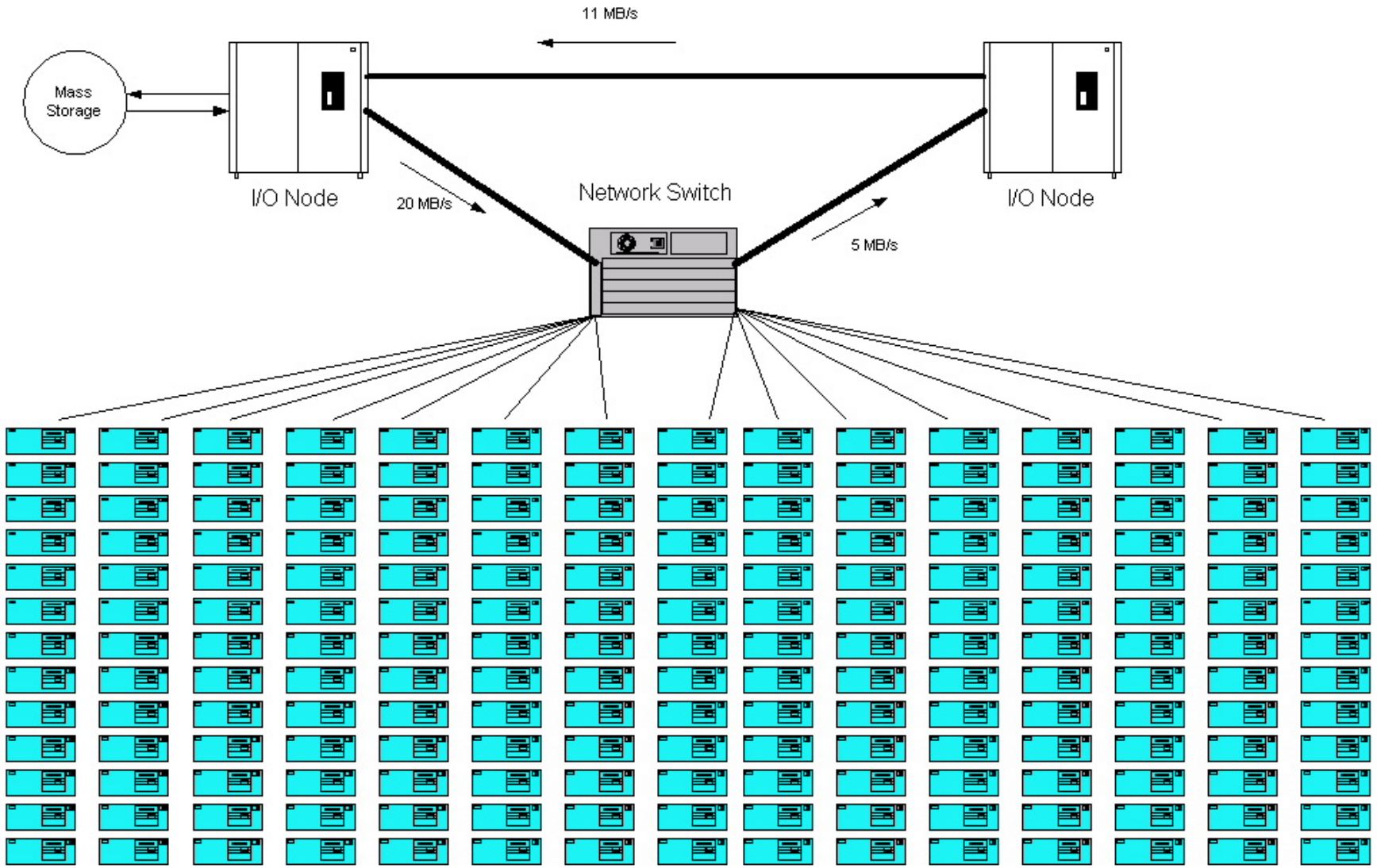
Design/Model

- **Hardware**
  - **Choose the most cost-effective CPU's for the compute-intensive computing**
  - **This is currently the dual-Pentium architecture**
  - **Network is fast and gigabit ethernet, with all machines being connected to a single or at most two large switches**
  - **A large I/O system to handle the buffering of data to/from mass storage and to provide a place to split the data into physics datasets**

Simple Model



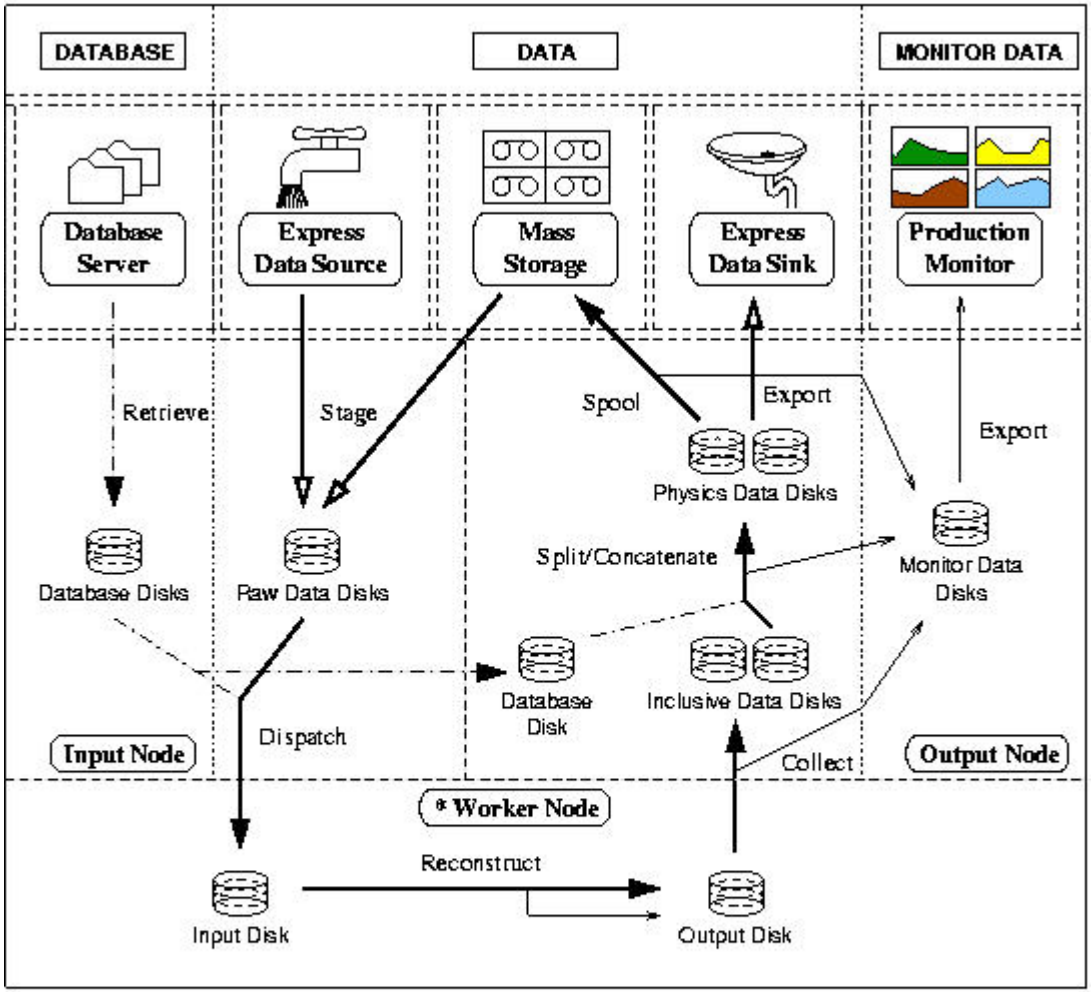
Run II CDF PC Farm

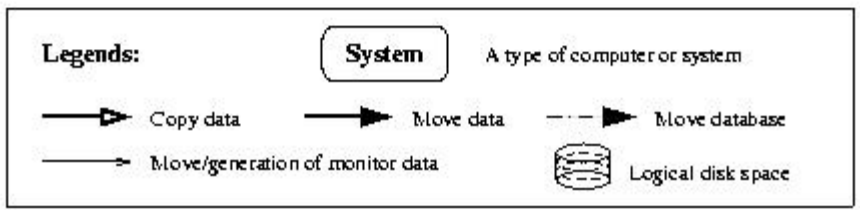


Software Model

- **Software consists of independent modules**
  - **Well defined interfaces**
  - **Common bookkeeping**
  - **Standardized error handling**
- **Choices**
  - **Python**
  - **MySQL database (internal database)**
  - **FBS (Farms Batch System) ***
  - **FIPC (Farms Interprocessor Communication) ***
  - **CDF Data Handling Software ***
  - *** Discussed in other CHEP talks**

Conceptual Model of Run 2 Production System



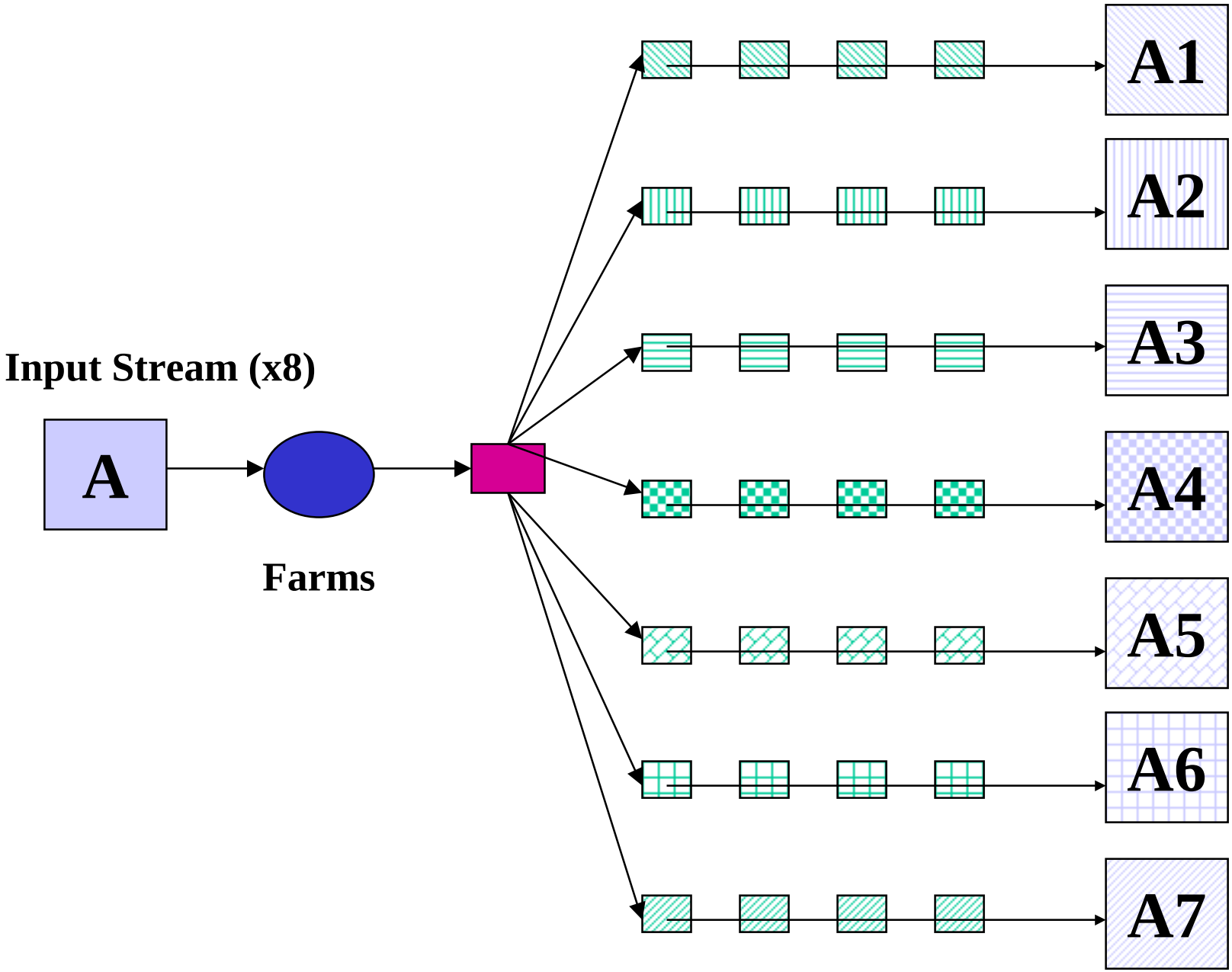


List of Software Modules

- **Coordinator**
- **Stager**
- **Dispatcher**
- **Reconstructor**
- **Collector**
- **Splitter**
- **Spooler**
- **Exporter**
- **Database Retriever**
- **Disk Manager**
- **Tape Manager**
- **Worker Node Manager**
- **Bookkeeper**
- **Messenger**

Physics Analysis Requirements and Impact

- **Raw Data Files come in ~8 flavors, or streams**
  - **1 Gbyte input files**
- **Reconstruction produces inclusive summary files**
  - **250 Mbyte output files**
- **Output Files must be split into ~8 physics datasets per input stream**
  - **Target 1 Gbyte files**
  - **About 20% overlap**
- **Leads to a complicated splitting/concatenation problem, as input and output streams range from tiny (<few percent) to quite large (10's of percent)**


Input Stream (x8)
A
Farms
A1
A2
A3
A4
A5
A6
A7

Prototype Farm/Mock Data Challenge 1

- **A small prototype farm has been used to test software, study hardware performance, and provide small but significant CPU resources**
- **4 I/O + 14 worker PII/400 dual Linux PC's**
- **Gigabit ethernet (I/O nodes/prototype farm)**
- **100 Mbit ethernet (worker nodes and SGI)**
- **Useful and necessary step before scaling up to larger farms**
- **Expected data rates were achieved**

  (20 Mbyte/s aggregate data transfer)

Prototype Farm

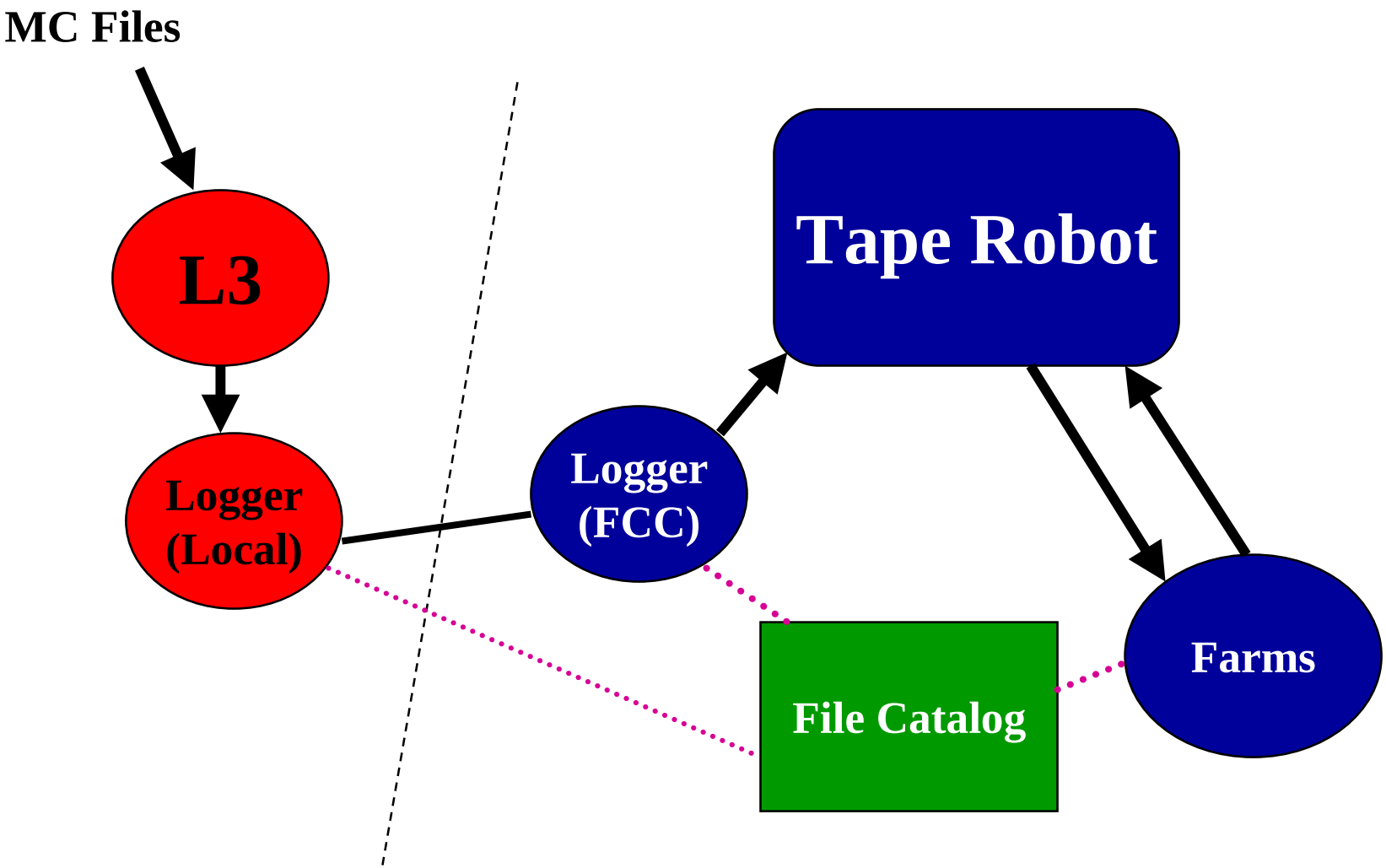
Mock Data Challenge 1 (CDF)

- **Primarily a connectivity test, with the following components:**
  - **Level 3 (Monte Carlo input)**
  - **Data Logger**
  - **Tape robot/mass storage system**
  - **File catalog (database)**
  - **Production executable (with all detector components and reconstruction)**
  - **Farm I/O + worker nodes**
  - **CDF Data Handling System**

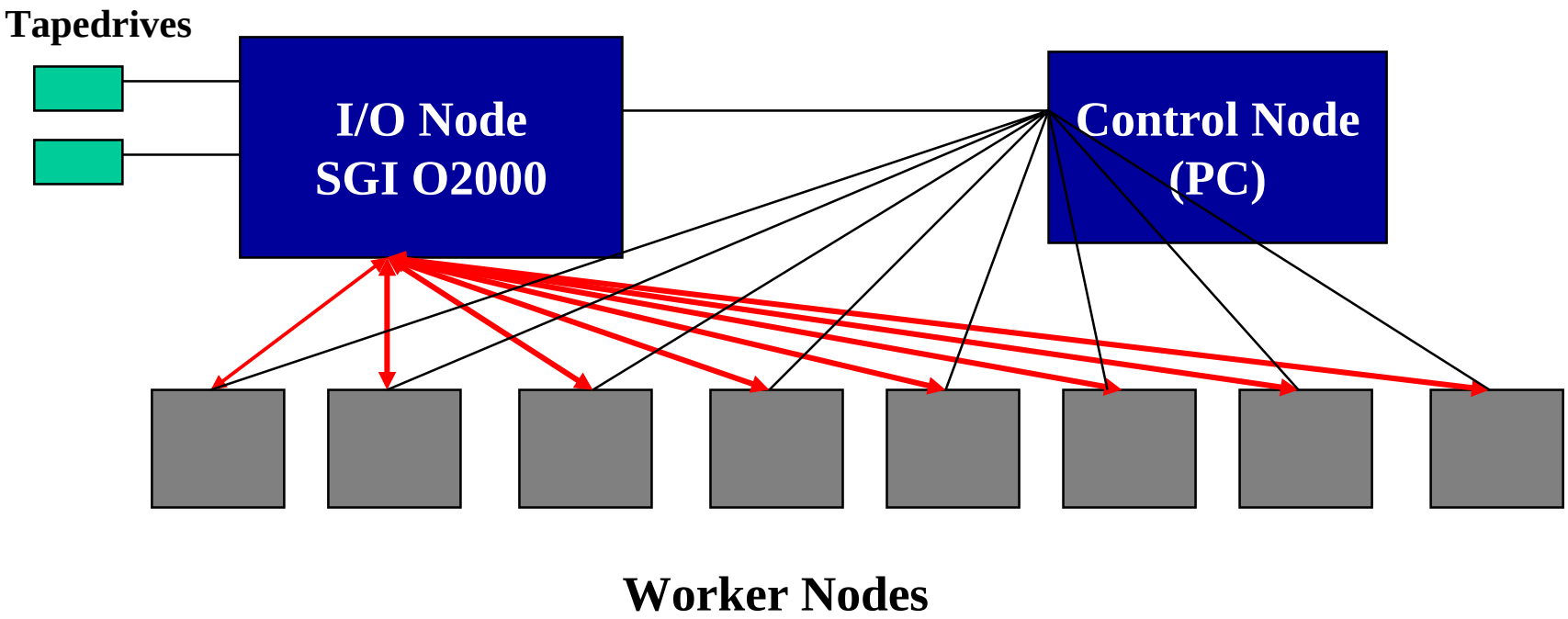
Mock Data Challenge 1

- **Monte Carlo events**
  - **2 input streams**
  - **6 output streams**
  - **1 Gbyte files, trigger bits set in L3 for splitting events**
- **Data Flow**
  - **Approx. 100 Gbyte of data was generated**
  - **Processed through L3 using L3 executable**
  - **Logged and processed on the farms through the full offline reconstruction package**

MDC1 - Data Flow



MDC1 Farms



Future

- **50 new PC's are in place (PIII/500 duals)**
  (Acquisition and testing was interesting)
- **New I/O node is being acquired (SGI O2200)**
- **Plan to integrate I/O systems, Cisco 6509 switch, 50 PC's**
- **Prepare for rate test and MDC2 (April/May 2000)**
- **Use same system for CDF Engineering Run in Fall 2000**


EXIT

Future (cont.)

- **Run II begins March 1, 2001**
- **Farms must grow to accommodate the expected data rate**
- **All PC's will be purchased as late as possible**
  - **PC's will increase from 50 to 150 or more**
  - **I/O systems should be adequate for full Run II rate**
  - **Switch should have sufficient capacity**

Future (software and process)

- **Software must be completed, debugged, and made user-friendly and supportable.**
- **CDF experimenters will monitor the farm.**
- **CDF farm experts will debug, tune and watch over the farm.**
- **Looking for a smooth, easy to run system**
  - **Must run for many years**
  - **Would like it to run with little intervention**

Far Future/possible scaling

- **System should scale beyond design.**
- **Can add I/O capacity by increasing disk storage, tapedrives, CPU speed, adding Gbit ethernet or 10 Gbit ethernet when available.**
- **More PC's can be purchased.**
- **The switch has capacity.**
- **If all else fails the entire system can simply be cloned, but overall control and database access is a potential issue in this case.**

Conclusion

- **The CDF farm is rapidly coming together:**
  - **Hardware**
  - **Software design and implementation**
- **Capacity for large-scale tests is almost in place.**
- **Plans for full Run II rate are firm.**
- **Upgrade path to go beyond Run II nominal rates exists.**